UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

ABDI NAZEMIAN, et al.,
Plaintiffs,
v.
NVIDIA CORPORATION,
Defendant.

Case No. 24-cv-01454-JST (SK)

ORDER REGARDING DISCOVERY LETTER BRIEF

Regarding Docket No. 162

Now before this Court is a dispute about the scope of discovery based on Plaintiffs' requests. For the reasons set forth below, the Court GRANTS IN PART and DENIES IN PART the request by Plaintiffs.

A. Background

Plaintiffs Abdi Nazemian, Brian Keene, and Stewart O'Nan (collectively, "Plaintiffs"), on behalf of themselves and all others similarly situated, allege that Defendant NVIDIA Corporation infringed their copyrighted books when it used databases containing that copyrighted material to train its large language models. (Dkt. No. 1.) Plaintiffs allege that Defendant announced the availability of four large language models in the "NeMo Megatron series": NeMo Megatron-GPT 1.3B, NeMo MegatronGPT 5B, NeMo Megatron-GPT 20B, and NeMo Megatron-T5 3B. (*Id.*) Defendant publicly disclosed that the NeMo Megatron large language model series was trained on "'The Pile' dataset prepared by EleutherAI." (*Id.*) The Pile dataset includes materials from "Books3." (*Id.*) Plaintiffs allege that their copyrighted books are included in Books3 and therefore in The Pile. (*Id.*) Therefore, Plaintiffs allege that Defendant trained its NeMo Megatron models on Plaintiffs' copyrighted books. (*Id.*) Plaintiffs allege direct copyright infringement under 17 U.S.C. § 501. They purport to represent a class of Plaintiffs from March 8, 2021 to the present, but they intend to amend the class period if they learn that infringement occurred earlier.

United States District Court
Northern District of California

Plaintiffs seek to represent the following class:

> All persons or entities domiciled in the United States that own a United States copyright in any work that was used as training data for the NeMo Megatron large language models during the Class Period.

(*Id.*)

A short and overly simplified explanation of the process of training a large language model is necessary to understand the dispute. There are datasets publicly available, and these datasets include large quantities of copyrighted material, used without consent of the owner of the copyrighted material. Companies such as Defendant then can use those datasets to "train" a large language model to learn patterns of speech, structure, and grammar. The large language model then learns to predict patterns and then can "understand, generate, and manipulate human language."[1]

B. Analysis

Plaintiffs seek discovery both about the datasets or libraries that Defendant used to train its large language models and the specific large language models that trained on those datasets. Defendant argues that any discovery should be limited to the large language models that trained on The Pile and that any discovery should also be limited to the four large language models identified in the Complaint.

Plaintiffs argue that, because they seek to represent a class of plaintiffs who own any copyrighted material from any source, they are entitled to discovery on all sources and for any large language models in the Nemo Megatron family that trained on those datasets. In other words, because Plaintiffs do not define the term "family" and do not limit their discovery requests to the four large language models identified in the Complaint, Plaintiffs seek information about any use of any dataset by Defendant.

With respect to the datasets, the Court finds that limiting the discovery to the dataset that Plaintiffs know that Defendant used and that contains Plaintiffs' copyrighted books is appropriate. Plaintiffs argue that they are hampered by lack of knowledge because Defendant is the only entity

[1] For example, this explanation was generated by artificial intelligence in response to a question submitted by the Court.

United States District Court
Northern District of California

that has access to information about its own activities. This is an inherent problem in any litigation: a plaintiff might not be able to discover wrongdoing by a defendant. That lack of knowledge does not justify a discovery request without bounds. Here, Plaintiffs know that Defendant used The Pile, which contains Plaintiffs' copyrighted books, to train its large language models. Discovery will thus be limited to training on The Pile.

With respect to the limitation on specific large language models, the Court finds that Plaintiffs are entitled to discover whether Defendant trained other large language models on The Pile. Here, Defendant's proposed limitation to a "family" of large language models is a meaningless distinction because Defendant has not defined "family." For this reason, the Court will not place an artificial restriction on the large language models that are the subject of Plaintiffs' search. They may seek information about large language models that use information from The Pile dataset by seeking information about other large language models in the Nemo Megatron family that exist and that trained on The Pile dataset.

Thus, the Court DENIES the motion to compel discovery regarding the use of datasets other than The Pile but GRANTS the motion to compel discovery regarding the use of The Pile to train datasets in the Nemo Megatron family beyond the four large language models specifically named in the Complaint.

IT IS SO ORDERED.

Dated: August 21, 2025

Sallie Kim

SALLIE KIM
United States Magistrate Judge